Análisis de calidad y recuperación de información en foros de discusión

Nadina Martinez Carod, Valeria Zoratto, Gabriela Aranda, Alejandra Cechich,
Agustín Chiarotto, Carina Noda, Mauro Sagripanti

Grupo de Investigación en Ingeniería de Software del Comahue (GIISCo) http://giisco.uncoma.edu.ar
Facultad de Informática. Universidad Nacional del Comahue -Buenos Aires 1400, (8300) Neuquén
Contacto: {nadina.martinez,vzoratto, gabriela.aranda, valeria.zoratto}@fi.uncoma.edu.ar

RESUMEN

Actualmente las organizaciones dedican mucho esfuerzo a resolver problemas utilizando estrategias ya probadas, lo que requiere brindar especial importancia a los sistemas de Information Retrieval (IR). Ante un problema que surge, lo primero que se analiza es si ha sido resuelto antes, bajo qué situaciones y en qué contexto se dieron problemas similares. Es aquí donde cobran gran importancia los sitios web donde se propician discusiones e intercambio de ideas sobre problemas comunes. Herramientas colaborativas como los blogs, wikis y foros de discusión hacen posible el reúso de conocimiento disponible en la Web. Entre ellas, nuestro proyecto se enfoca en los foros de discusión como herramienta fundamental, ya que contiene una base de conocimiento lo suficientemente completa para ser reutilizada. La enorme cantidad de información existente en los foros, sumado a la simplicidad de uso de los mismos, hacen necesario definir estrategias para clasificar las soluciones ya probadas, fundamentales en los sistemas IR.
De esta manera el principal objetivo de este proyecto es la creación de una herramienta que realice una clasificación automática de la información contenida en foros de discusión, utilizando la información existente en los hilos de discusión, que luego de ser analizada, será procesada, reutilizada y clasificada para solucionar problemas específicos recurrentes que fueran surgiendo.

CONTEXTO

Nuestro proyecto se enmarca en el Programa "Desarrollo de Software Basado en Reúso – Parte II", de la Universidad Nacional del Comahue, a realizarse en el período 2017-2020, el cual extiende al Programa "Desarrollo de Software Basado en Reúso" realizado en el período 2013-2016. Dicho Programa está basado en el proceso sistemático de detectar, seleccionar, organizar, presentar y utilizar la información del capital intelectual de las organizaciones, con el fin de brindar mayor competitividad a las mismas, aplicando el conocimiento en el problema objeto para resolver. Dicho Programa está compuesto por tres subproyectos los cuales coinciden en el tratamiento del desarrollo de software basado en reúso, pero desde aspectos diferentes: orientado a dominios, orientado a servicios y orientado a foros de discusión. El proyecto actual, denominado "Reúso de Conocimientos en Foros de Discusión – Parte II", continúa la línea de investigación enfocada en la recuperación de información y de conocimiento disponible en foros de discusión técnicos.

1. INTRODUCCIÓN

Los foros de discusión son utilizados para consultar, discutir e intercambiar conocimiento en la Web, pero su principal característica es que facilita la comunicación de personas distribuidas geográficamente. Se puede describir un foro como un espacio virtual de una comunidad de usuarios con temas de interés común. De esta manera

encontramos foros con temas variados como viajes, cocina, historia, software entre otros, sin embargo, en nuestro proyecto nos interesan en particular los foros sobre temas aplicados a las ciencias de la computación.

Dentro de un foro, los participantes no necesitan encontrarse conectados al mismo tiempo (herramienta asíncrona), incluso en la gran mayoría de las veces los participantes no se conocen personalmente, pero sí a través de sus nombres, alias o avatares (representaciones gráficas que se asocian a usuarios para identificarse). Algunos de estos foros, como Stackoverflow[1], se destacan por ofrecer entre sus soluciones fragmentos de código, así como información valiosa de los usuarios (por ejemplo, la reputación que se construye a partir de su participación y de votos de los otros usuarios). Otros, como lawebdelprogramador[2] arman una valoración del usuario de manera automática a partir de información del usuario que van obteniendo a lo largo de los mensajes que estos escriben, como preguntas iniciales, respuestas dadas y puntuación de sus miembros. Esto implica que, al trabajar con foros, como cada uno tiene una estructura diferente, se complejiza la tarea de recopilación de información y del análisis a realizar sobre ella. Estas características hacen de los foros de discusión no sólo una herramienta fundamental que facilita el trabajo colaborativo y distribuido, sino también una fuente esencial de conocimiento que puede ser reutilizado.

De acuerdo al permiso dado a los participantes para ver o comentar dentro del foro, se pueden distinguir diferentes tipos de foros: en los públicos todos los participantes pueden comunicarse o leer mensajes escritos por el resto sin necesidad de registrarse; en cambio en los foros protegidos es necesario registrarse para luego poder enviar mensajes. Por último, en los foros privados se exigen ciertas restricciones para participar y utilizar la información.

[1] https://stackoverflow.com/

[2] https://www.lawebdelprogramador.com/

Como el proyecto se enfoca en trabajar con la información existente en foros de discusión disponibles para cualquier usuario, sin necesidad de tener una registración en los mismos; los foros elegidos para nuestras evaluaciones son públicos, donde la totalidad de los mensajes queda disponible para ser consultada en problemas similares.

Respecto a la estructura de un foro de discusión, el mismo contiene una colección de hilos. Un hilo se genera a partir de una pregunta inicial sobre un problema específico. A partir de ese momento, los usuarios de la comunidad podrán debatir en función del tema y del problema enunciado. Luego, para obtener conocimiento proveniente de los hilos de discusión se utilizan diferentes técnicas y estrategias para establecer cuáles de las posibles soluciones obtenidas de los foros pueden ser relevantes para consultas sobre problemas similares.

El proyecto realiza por un lado el tratamiento del texto contenido en los hilos dentro del foro. Luego, la búsqueda transita por caminos independientes, pero con un objetivo en común: complementar el conjunto de estrategias de clasificación de hilos para definir un orden de prioridad. Para ello se ha continuado con el enfoque de Elsas & Carbonelli [6], el cual utiliza la estructura de los hilos separando la pregunta inicial del resto del hilo, tratándolo con la dupla <pregunta, hilo>. Una estructura similar se logra también mediante el uso de patrones Feng & Shaw [7], y de otros tratamientos como el de Cong & Wang [5] que separan preguntas de respuestas con un robot reconocedor de campos condicionales. Otra orientación que se ha analizado es la de clasificar o estructurar temas mediante jerarquías, como el enfoque de Nicoletti [17] o el de Helic et al. [11]. También se analizó la concordancia con Liu et al. [12], que busca la satisfacción de las respuestas sobre una pregunta realizada, o el enfoque que presenta Bathia [4] a partir del análisis de expertitud de

los usuarios, detectando niveles de conocimiento de los comentarios, para clasificar con mayor valoración los hilos en los cuales intervienen personas expertas o con altos conocimientos.

Otro camino analizado es mediante algoritmos de lenguaje natural para clasificar tipos de fragmentos dentro de los hilos de discusión utilizando técnicas de aprendizaje automático, como la propuesta por Tigelar et al. [19]. Dentro de este camino se encuentran también técnicas de minería de datos, para la búsqueda de patrones y reglas significativas, haciendo énfasis en la minería web (minería de datos sobre la web), tanto en procesamiento de lenguaje natural como opinion mining, en conjunto con la aplicación de *sentiment analysis*. Este estudio está focalizado en la búsqueda de patrones reconocedores de una determinada respuesta sobre la pregunta en un hilo, para determinar si tuvo aceptación o rechazo, y en qué escala. De esta manera se puede detectar aquellos hilos donde la respuesta fue satisfactoria y el grado de satisfacción del mismo.
Teniendo en cuenta las distintas direcciones mencionadas, se realizaron actividades, con resultados favorables en su mayoría, permitiendo la extensión de algunas líneas de avance y la elaboración de nuevas líneas a favor del objetivo del proyecto.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación principal está enmarcada en un Programa de investigación de la Universidad Nacional del Comahue, llamado "Desarrollo de Software Basado en Reúso – Parte II", dentro del período comprendido entre los años 2017 y 2020. El programa tiene 3 líneas de investigación, relacionadas; reúso orientado a dominios, reúso orientado a servicios y la nuestra, cuya denominación es "Reúso de Conocimientos en Foros de Discusión – Parte II". El programa está desarrollado por el Grupo GIISCo de la Facultad de Informática, y su objetivo es continuar con las investigaciones iniciadas en el programa inicial, período 2013-2016 denominado "Desarrollo de Software Basado en Reúso", perteneciente a la misma universidad.
El Grupo GIISCo está conformado por docentes de la Facultad de Informática de la Universidad Nacional del Comahue, no sólo del área de Ingeniería de Software, sino también incluye docentes con otras ópticas establecidas por las áreas de Programación y Teoría de la Computación. Esto, sumado a la colaboración de la Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires, le brinda un marco heterogéneo propicio para la investigación de una amplia gama de soluciones.

3. RESULTADOS OBTENIDOS/ESPERADOS

A partir del objetivo de nuestro proyecto, el cual es brindar asistencia inteligente en la recomendación de hilos de discusión para buscar soluciones a problemas recurrentes presentados en foros de discusión técnicos, podemos citar resultados obtenidos hasta el momento y mencionar lo que se espera a partir del presente.
Los resultados obtenidos sobre nuestra línea de investigación parten de un modelo de calidad para foros de discusión en base a modelos de calidad de datos e información en la Web y estándares para la calidad de datos software, En base a de dicho modelo se determinaron métricas, que fueron validadas mediante encuestas [1], posteriormente se avanzó en el diseño e implementación de una herramienta para la recuperación de información [14], aplicando un grupo de métricas de calidad [21]. Para poder manipular la información de foros, se trabajó en el análisis y tratamiento de textos, con herramientas como Lucene[3], lo que dio lugar a tratamientos de recuperación de información para lenguaje específico de Java, [2, 22]. En esta línea se continuó con la utilización de

[3] https://lucene.apache.org

sinónimos, mediante el uso de la base de datos léxica en inglés WordNet[4] [22], mediante un analizador morfológico como Stanford POS Tagger[5] [3]. La tercera línea que se siguió fue la clasificación del conocimiento de los usuarios participantes de los hilos de discusión. Esta línea se basó en las ópticas de Lui & Baldwin [13], la de Bathia & Mitra [4] y la de Hecking et al. [10], bajo las cuales se diseñó una estrategia para determinar una jerarquía de roles determinados por el nivel de conocimientos de los participantes en los hilos de acuerdo a los posts realizados [15].

Considerando los resultados obtenidos hasta el momento como punto de partida, los resultados que se esperan obtener continúan en distintas direcciones, pero con nuevas alternativas. Por un lado, se continúa trabajando en la tarea de clasificación de tipos de mensajes para poder destacar las preguntas de las respuestas, informando la retroalimentación positiva y negativa, mediante la utilización de técnicas de procesamiento de lenguaje natural (PNL), utilizando técnicas de Data Mining, a partir de la propuesta de Liu [12], tanto en modelos de aprendizaje supervisados como no supervisados presentadas por Witten [20], además, siguiendo la línea de Shoji et al.[18], se está analizando al usuario que realiza la pregunta, teniendo en cuenta su historia dentro de la comunidad, para detectar posibles patrones de comportamientos Esta línea de investigación se sigue desarrollando en una tesis de doctorado en la cual se evalúa en conjunto: el rol del post, respuestas de calidad, expertitud del usuario que responde y satisfacción del usuario que pregunta.
En otra dirección se continúa trabajando en estrategias para la obtención de roles, aplicando un análisis de redes sociales en patrones de respuestas de los participantes que han escrito un post dentro del mismo hilo, propuesto por Fisher et al. [8]. Siguiendo con la línea de análisis de los usuarios participantes, se está realizando una tesis para buscar una estrategia para predecir la satisfacción de los usuarios que preguntan respecto a las soluciones propuestas por otros participantes.

[4] https://wordnet.princeton.edu/

[5] https://nlp.stanford.edu/software/tagger.shtml

4. FORMACIÓN DE RECURSOS HUMANOS

Debido a que la naturaleza del proyecto es multidisciplinaria, dentro de sus integrantes hay docentes de diferentes áreas. En particular en las de Ingeniería en Sistemas, Programación, y Teoría de la Computación. Las personas que forman parte del proyecto, tanto como colaboradores, asesores o integrantes son:

- 2 profesores adjuntos con dedicación exclusiva.
- 1 docente investigador con beca del CONICET.
- 2 docentes con dedicación simple.
- 1 profesora adjunta, asesora de la UNCo, con dedicación exclusiva.
- 1 profesora adjunta, asesora externa de la UNCPBA.
- 5 estudiantes de la carrera de Licenciatura en Ciencias de la Computación.

Los integrantes del proyecto que pertenecen al Departamento de Programación son dos docentes con dedicación exclusiva, doctores en Informática; el otro docente es asistente simple con beca de CONICET para hacer el doctorado, un docente investigador con dedicación simple. El otro docente investigador con dedicación simple pertenece al Departamento de Ingeniería de Sistemas.
La docente colaboradora de la UNCo que pertenece al área de Teoría de la Computación, posee un Diploma de Estudios Avanzados y se encuentra realizando la Maestría en Ciencias de la Computación de la Universidad Nacional del Comahue.
También integra el proyecto una asesora externa que pertenece a la Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), con experiencia en

Sistemas de Recomendación y Recuperación de Información, y pertenece al Instituto Superior de Ingeniería de Software (ISISTAN). Dicha docente posee el título de Doctora en Ciencias de la Computación.

Tres estudiantes de Licenciatura en Ciencias de la Computación desarrollaron sus tesis dentro del proyecto, uno de cuales está en etapa de defensa de la misma. Además, se han sumado cuatro estudiantes, los cuales están realizando sus tesis. De esta manera, se van incorporando actividades para extender líneas de investigación al proyecto inicial con nuevos enfoques.

5. BIBLIOGRAFÍA

[1] G. Aranda, N. Martínez Carod, P. Faraci, A. Cechich. Hacia un framework de evaluación de calidad de información en foros de discusión técnicos. ASSE 2013,

[2] G. Aranda, N. Martínez-Carod, S. Roger, P. Faraci, and A. Cechich. Una herramienta para el análisis de hilos de discusión técnicos. In CACIC 2014, pages 803 - 812, 2014.

[3] G. Aranda, V. Zoratto, N. Martínez Carod, Sandra Roger, F. Otermín, A. Cechich. Clasificación de contenido de hilos de discusión mediante análisis sintáctico y morfológico. CICCSI 2018.

[4] S. Bhatia and P. Mitra. Classifying user messages for managing web forum data. In Z. G. Ives and Y. Velegrakis, editors, WebDB , pages 13-18, 2012

[5] G.Cong, L. Wang, C. Lin, Y. Song, and Yueheng (2008). Finding question-answer pairs from online forums. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). Association for Computing Machinery, 467–474.

[6] J.L Elsas and J. G Carbonell, "It pays to be picky: an evaluation of thread retrieval in online forums", in Proceedings of the 32nd international ACM SIGIR (2009), pp. 714—715.

[7] D. Feng, E. Shaw, J.Kim & E.Hovy (2006). An intelligent discussion-bot for answering student queries in threaded discussions. In Proceedings of the 11th international conference on Intelligent user interfaces (IUI '06). Association for Computing Machinery, 171–177.

[8] D. Fisher, M. Smith, and H. T. Welser, You are who you talk to: Detecting roles in usenet newsgroups, in Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), vol. 3, pp. 59b59b, IEEE, 2006.

[9] A. Gangemi, R. Navigli, P. Velardi. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet, In Proc. of ODBASE 2003, Catania, Sicily (Italy), 2003, pp. 820–838.

[10] T. Hecking, I. Chounta, and H. U. Hoppe. Investigating social and semantic user roles in MOOC discussifon forums. In LAK, pages 198-207. ACM, 2016

[11] D. Helic, N. Scerbakov (2003), "Reusing Discussion Forums as Learning Resources in WBT Systems".

[12] B. Liu. Web Data Mining. Exloring Hyperlinks, Contents, and Usage Data. Springer. 2008

[13] M. Lui and T. Baldwin. Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet. In Proceedings of Australasian Language Technology Association Workshop, pages 49-57, 2010.

[14] N.Martínez Carod, G. Aranda. Análisis de la información presente en foros de discusión técnicos. In CACIC 2013, pp. 847- 856, 2013.

[15] N.Martínez Carod, G. Aranda. Valeria Zoratto, Christian Murray. Una propuesta para clasificación de roles de usuarios en foros de discusión técnicos. In CACIC 2019, pp. 836- 845, 2019.

[16] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An online lexical database. Int. J. Lexicograph.pp. 235–244.

[17] M. Nicoletti, S. Schiafino, and D. Godoy. Mining interests for user profiling in electronic conversations. Expert Syst. Appl., Feb. 2013.

[18] Y. Shoji, S. Fujita, A. Tajima, and K. Tanaka. Who stays longer in community qa media?-user behavior analysis in cqa. In International Conference on Social Informatics, pages 33 48. Springer, 2015. .

[19] A. Tigelaar, R. Op Den Akker and D. Hiemstra, Automatic summarisation of discussion fora, Natural Language Engineering, ISSN 1469-8110, Vol 16, Issue 02, pp. 161-192, 2010.

[20] I. Witten, E. Frank and M. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier. 2011

[21] V. Zoratto, G. Aranda, S. Roger, A. Cechich, Análisis de estrategias para clasificar contenidos en foros de discusión: Un caso de estudio ASSE 2015, pp. 176-190.

[22] V. Zoratto, G. Aranda, S. Roger, A. Cechich, Analyzing Discussion Forums ThreadsAbout Java Programming Language Usage, Electronic Journal of SADIO, 2016.